Spring 5-1-2018

The Identification of Functional CLL-Associated SNPs Through PRO-seq and Transcription Factor Binding Site Analysis

Mary Accurso
mary.accurso@uconn.edu

The Identification of Functional CLL-Associated SNPs Through PRO-seq and Transcription Factor Binding Site Analysis

Mary Accurso
University of Connecticut
Department of Molecular and Cell Biology
Honors Thesis – May 2018

Thesis Advisor: Dr. Leighton Core, Ph.D.
Honors Advisor: Dr. Leighton Core, Ph.D.

Table of Contents

I. Abstract

Gene expression is essential for every cellular process in our bodies. Improper regulation of gene expression can cause a variety of serious conditions such as cancer. Single nucleotide polymorphisms (SNPs) are single base pairs that vary based on the individual. Some SNPs are located in regions of the genome essential for regulation of gene expression, while other SNPs are located in non-coding regions of the genome where they might cause some effect that is currently unknown.

Chronic lymphocytic leukemia (CLL) mainly affects adult populations and has a high familial risk component compared to other cancers. Several SNPs have been associated with the disease, however many are located in non-coding areas of the genome and their function is not yet known. Our hypothesis is that these SNPs function as regulatory elements that affect the transcription of nearby genes. It is believed that not a single SNP, but multiple SNPs work together to generate the CLL disease phenotype. The objective of this research is to locate previously unidentified SNP(s) that are in linkage disequilibrium (LD) with known CLL-associated SNPs, and to identify their function.

Since many CLL-associated SNPs are located in non-coding areas of the genome, functional genomics studies must be done to determine if they are located in areas with some functional activity. Towards this end, precision run-on sequencing (PRO-seq) was performed on 5 normal and 18 CLL subject samples. Discriminatory regulatory-element detection from GRO-seq (dREG) allowed us to identify transcription regulatory elements (TREs) and to quantify their activity from the PRO-seq data (Danko et al., 2015). TREs with a genotypic specific change in activity became candidates for further analysis. Primers were designed around TREs in regions

close to three known CLL-associated SNPs that were also in areas of LD. PCR and Sanger Sequencing were performed on these regions to identify other SNPs in the region.

Additional SNPs were identified during the sequencing of the candidate TREs, and transcription factor (TF) binding site analysis was performed. A SNP on chr16, position 85,934,116, was identified in a TRE in LD with the known SNP rs305088. TF binding site analysis showed that this SNP acts as a binding site for multiple TFs. As the SNP is in an area of LD with rs305088, the two SNPs are more likely to segregate together during meiosis. This suggests that multiple SNPs segregate together during meiosis and together contribute to the genetic inheritance of CLL, possibly through the regulation of TF binding sites. The linked segregation and inheritance of this SNP may contribute to the genetic component and the progression of CLL by preventing or enhancing TF binding to the region, thereby preventing normal function or enhancing a new, abnormal function.

II. Introduction

Chronic lymphocytic leukemia (CLL) is the most commonly diagnosed type of leukemia in adults in western countries and is responsible for about a quarter of all cases of diagnosed leukemia (Ghia et al., 2007). The disease is about twice as common in men as it is in women and about 15,000 new cases of CLL are diagnosed in the United States every year (Ghia et al., 2007). CLL is characterized by the accumulation of B lymphocytes in the peripheral blood, bone marrow, lymph nodes, and spleen. Despite various treatment options, the disease remains incurable (Ghia et al., 2007).

The etiology of CLL also remains unknown (Kalil, 1999). Environmental factors such as radiation, chemicals, and drugs have shown no evidence of causing the disease (Kalil, 1999). However, some research has suggested that there may be an inheritable genetic component to the

disease (Kalil, 1999). One study showed that first-degree relatives to patients with CLL have an ~8.5 increased risk compared to individuals with no relatives with CLL (Goldin et al., 2009).

CLL cells can contain several genomic aberrations, but a unique and causative genomic aberration has yet to be discovered (Ghia et al., 2007). Although a single genomic locus associated with the disease has not been identified, numerous genome-wide association (GWA) studies have identified multiple susceptibility loci for CLL. These GWA studies have genotyped hundreds of thousands of SNPs in large cohorts of patients with CLL. One study identified six SNP loci that were estimated to account for ~3% of the familial risk of CLL (Di Bernardo et al., 2008). A second, follow-up study identified four additional risk loci and estimated that the ten total susceptibility loci account for ~10% of the familial risk of CLL (Crowther-Swanepoel et al., 2010). A third study in 2011 focused on familial CLL cases in an attempt to identify SNPs specific to familial CLL and identified an additional risk locus, bringing the total to eleven identified risk loci (Slager et al., 2011). It has been estimated that SNPs could account for up to ~46% of the familial risk associated with CLL (Berndt et al., 2013).

Although at least 35 GWAS-discovered SNPs have already been identified, they only account for a fraction of the genetic susceptibility of CLL, ~17% (Slager et al., 2013). Six of these SNPs were chosen in collaboration with Dr. Jennifer Brown's lab at the Dana-Farber Cancer Institute, however, three of these SNPs appear to be in non-transcribed regions of the genome so we are currently only investigating the three SNPs in actively transcribed regions. The three SNPs we chose were: rs305088, located on chromosome 16 and a SNP proxy for rs305061 (Crowther-Swanepoel et al., 2010); rs4777184, located on chromosome 15 and a SNP proxy for rs7176508 (Di Bernardo et al., 2008); and rs674313, located on chromosome 6 and identified in 2011 (Slager et al., 2011).

When SNPs are located in coding regions it is relatively easy to determine their function. One can study the effect of a SNP on the amino acid sequence and subsequent structure and function of the protein. However, when SNPs are located in non-coding regions of the genome it is harder to determine their function and therefore functionality tends to remain unknown. In order to study the role of SNPs located in non-coding regions, functional genomics can be used to determine if these SNPs are in regions of the genome with some functional activity. Precision run-on and sequencing (PRO-seq) allows us to do this by identifying transcription regulatory elements (TREs) and quantifying TRE activity.

PRO-seq allows us to sequence all actively transcribed nascent RNA which allows us to study the transcriptome at a very high resolution (Kwak et al., 2013). Endogenous nucleotides are washed away which pauses the RNA polymerase II until biotin-labeled nucleotides are introduced. RNA polymerase II is only able to incorporate one or two biotin-labeled nucleotides onto the end of the nascent RNA and then streptavidin beads are used to extract only the biotin-labeled nascent RNA for sequencing. As a result, we are able to sequence all actively transcribed RNAs. After sequencing, discriminatory regulatory-element detection from GRO-seq (dREG) was used to identify TREs, which include promoters, enhancers, and insulators, and to quantitate the transcriptional levels of these TREs (Danko et al., 2015). dREG works by detecting divergent transcription, a characteristic pattern of active TREs (Core et al., 2008). If we find a genotypic specific transcription change in a TRE, then this region becomes a candidate for future study.

Gene expression is critical to every process that occurs in the body and major regulation of gene expression happens at the transcriptional level. SNPs can change the way that different regulatory mechanisms, including transcription factors (TFs), interact with the DNA which can affect transcription and could potentially lead to disease, including cancer. My research aims to

locate previously unidentified SNPs that are in linkage disequilibrium (LD) with GWAS-identified CLL-associated SNPs and to determine if these SNPs affect transcription by interfering with the binding of TFs to regions of the genome involved in the CLL phenotype and therefore might be a causal SNP of CLL. LD refers to the fact that alleles located close together on a chromosome have a greater chance of being inherited together because they are less likely to be separated during crossing over in meiosis. The SNPs that have been identified by GWAS studies are not necessarily causal SNPs - they might just be SNPs that are in LD with the causal SNP. The causal SNP will not only correspond to genotypic specific changes in TRE activity, it will also have some sort of functionality that can be linked to the disease phenotype.

III. Methods

a. Sequencing

Cellular nuclei were permeabilized and PRO-seq was performed on 5 normal and 19 CLL patient samples. dREG was used to call TREs and quantify TRE activity. As seen in Table 1, primers were designed around TREs in regions close to three known CLL-associated SNPs that had genotype-specific differences within the regions of LD - rs305088, rs4777184, and rs674313. Primers for the region surrounding rs674313 are not included because none of them were successful as will be discussed in the results section.

Primer Name	Sequence
rs305088U1_Fw	GAGGCTGACAGAGGAGAAATG
rs305088U1_Rv	GGTGAGGTGAGATCCTGAAAC
rs305088U2_Fw	GATTCTCACCCATCTCCCATTT
rs305088U2_Rv	CTGTTCATGCCTCCTCATAGTT
rs305088U3_Fw	GCCATAGGAATCTCACACAGAA
rs305088U3_Rv	TCACACCCAGAGAAAGGAAAC
rs305088U4_Fw	GTGATGGTGAGGTTGGGTAAG
rs305088U4_Rv	GACAAAGTCCAGGAAGGAAAGA

rs305088U5_Fw	CCTAGAGGTCAACAGCAACTG
rs305088U5_Rv	GAAAGCCCGCTCTGAAAGA
rs305088U6_Fw	TGAGTTACACGCATCTTCTTCTC
rs305088U6_Rv	CAGGAGGTCAGGACAATGAATC
rs305088U3_BegFw	CACAGGTAGGTGAGACAACTG
rs305088U3_BegRv	GAGAAGCCAGGTGAGAACAA
rs305088U3_MidFw	TCTGCCACCCTCGTCTT
rs4777184_Fw	CCCAAGGTCTCACAGCAATTA
rs4777184_Rv	TGGAAGAACCAAAGGGAGATG
rs4777184_FwSeq	CTTGTGACCCAGTCTCTACATT

Table 1: Primer names and sequences

Genomic DNA (gDNA) was extracted from 23 gDNA samples (we only had gDNA from 18 out of the 19 CLL patients samples that PRO-seq was performed on) and from two CLL cell lines – MEC1 and OSU-CLL. Polymerase chain reaction (PCR) was performed with the above primers to amplify the TRE regions in the 23 patient samples and cell lines. Typical PCR conditions were 30 cycles of touchdown PCR. Sanger sequencing was performed to locate other SNPs in these regions. The Sanger sequencing data is located under the "CLL Project" Inventory and respective chromosome template on Benchling. The primer sequences are in "The_Oligo_Database" on Benchling.

b. Transcription Factor Binding Site Analysis

The SNPs located using Sanger sequencing were analyzed using TOMTOM, a motif analysis tool that compares a query motif to a database of known TF motifs (Gupta et al., 2007). Sequences of 6bp upstream and downstream of the SNPs were analyzed for TF binding with the "risk" and the "normal" allele. The "Eukaryote DNA" and "JASPAR Vertebrates and UniPROBE Mouse" databases were used as the motif database to which the query motif was compared in all analyses except for the initial analysis of the new SNPs located at chr16:85,936,495 and chr16:85,936,497 in the rs305088_U3 TRE region. The initial motif analysis of these two SNPs

was analyzed using the "HUMAN (Homo sapiens) DNA" and "HOCOMOCO Human (v11 CORE)" databases. Differences in the TFs that bind to the sequence with and without the risk genotype at the SNP location were noted.

c. ChIP-seq Analysis

These differentially bound TFs were further studied using chromatin immunoprecipitation and sequencing (ChIP-seq) data from the Encyclopedia of DNA Elements (ENCODE) Consortium to assess TF binding to DNA with and without the risk SNP. As shown in Table 2, we analyzed 9 datasets from 8 TFs that had been identified using TOMTOM. Each dataset had 2 duplicates of raw data in FASTQ file format.

Antibody	Cell Line	Treatment	ChIP-seq Database	Replicate	Reads per Replicate
Egr1	K562	N/A	ENCODE	Rep1	18,848,323
	K562	N/A	ENCODE	Rep2	27,984,930
Tcf3	Gm12878	N/A	ENCODE	Rep1	6,868,500
	Gm12878	N/A	ENCODE	Rep2	15,242,225
Tcf12	Gm12878	N/A	ENCODE	Rep1	14,051,006
	Gm12878	N/A	ENCODE	Rep2	13,629,919
NFYA	Gm12878	N/A	ENCODE	Rep1	11,383,438
	Gm12878	N/A	ENCODE	Rep2	11,943,138
IRF1	K562	IFNa6h	ENCODE	Rep1	23,565,682
	K562	IFNa6h	ENCODE	Rep2	22,938,815
IRF1	K562	IFNg30	ENCODE	Rep1	18,645,852
	K562	IFNg30	ENCODE	Rep2	23,793,998
Spi1	K562	N/A	ENCODE	Rep1	13,887,722
	K562	N/A	ENCODE	Rep2	17,252,822
Yy1	K562	N/A	ENCODE	Rep1	10,803,175
	K562	N/A	ENCODE	Rep2	6,784,368
Tead4	K562	N/A	ENCODE	Rep1	27,988,989
	K562	N/A	ENCODE	Rep2	32,993,186

Table 2: ChIP-seq datasets used for analysis. Reads per replicate refers to the number of uniquely mapping reads.

The ChIP-seq data was mapped to hg38 using Bowtie (Langmead et al., 2009). Individual ChIP-seq replicates were analyzed using MACS2 to call narrow peaks (Zhang et al., 2008).

Bedtools intersect was used to find the high confidence peaks from the two replicates of each ChIP-seq data set (Quinlan et al., 2010). The summits of these peaks were identified and were then extended 50bp upstream and downstream to obtain uniform peak sizes of 101bp. The high confidence peaks were converted to FASTA files to be used by MAST. Individual MEME motif files were created for each TF identified by TOMTOM and used to run MAST with the high confidence peaks (Bailey & Gribskov, 1998).

Two MEME motif files were created for each TF – one for the TF binding site with the normal allele at the SNP location, and one for the TF binding site with the risk allele at the SNP location. Figure 1 shows an example of what a MEME motif file looks like as well as the position-weight matrix image of the TF and the genomic sequence with the risk allele at the SNP location. All of the information included in the MEME motif files is publicly available. The only thing that was manually changed was the letter-probability matrix.

Each column in the letter-probability matrix represents a nucleotide and the nucleotides are in alphabetical order, so from left to right the columns represent A, C, G, and then T. Each row represents a nucleotide position in the motif starting from position 1. Since each TF motif has a different number of nucleotides, the number of rows varies depending on the TF. The numbers in the matrix represent the probability of that nucleotide being present at a particular location in the motif. By changing the probabilities of the nucleotides at the SNP location, we can analyze the number of ChIP-seq peaks that match the motif when there is a 100% chance the risk allele is present at the SNP location, versus the number of ChIP-seq peaks that match the motif when there is a 100% chance the normal allele is present. The motif file in figure 1 is an example of when the TF binding site has the risk allele, G, at the SNP location. The SNP is located at position 6 of the

motif so the third column of row 6 of the matrix has a value of 1.000000 while the other nucleotides at position 6 have a value of 0.000000.

```
MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies
A 0.25 C 0.25 G 0.25 T 0.25

MOTIF MA0162.2 EGR1
letter-probability matrix: alength= 4 w= 14 nsites= 12256 E= 0
0.089589        0.467363        0.251550        0.191498
0.122879        0.558665        0.110884        0.207572
0.094648        0.493554        0.183584        0.228215
0.108926        0.851093        0.000000        0.039980
0.019011        0.944354        0.000000        0.036635
0.000000        0.000000        1.000000        0.000000
0.000000        1.000000        0.000000        0.000000
0.000000        0.967037        0.000000        0.032963
0.000000        0.828492        0.049853        0.121655
0.297977        0.682196        0.000000        0.019827
0.000000        0.975196        0.000000        0.024804
0.193211        0.000000        0.538430        0.268358
0.000000        0.803606        0.115862        0.080532
0.246899        0.456511        0.156087        0.140503
```

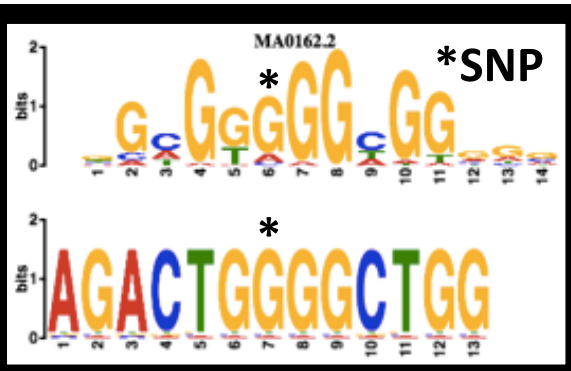


Figure 1: MEME motif file for EGR1 and position-weight matrix for EGR1 binding site.

The data was normalized by taking the coordinates of all detected TREs, finding the center of each TRE region and then extending the window 50bp upstream and downstream to get a 101bp region. Each TF motif was run against these TRE sequences using MAST. The number of observed sequences from the ChIP-seq database MAST results over the total number of sequences from the TRE database MAST results was used as the normalized data.

d. Cell Culture

For future CRISPR experiments which might be done to modify the SNP allele and observe the effect on transcription, experiments were performed to attempt to grow monoclonal MEC1 and OSU-CLL cells. A cell count of the respective cell line was performed using a hemocytometer and dilutions were made in order to get a concentration of 1 cell/100uL or 2 cells/100uL. Three types of media were used: 100% new RPMI media, 100% conditioned RPMI media, and a mixture of 50% new and 50% conditioned media. These dilutions were plated on a 96-well plate – 32 wells for each type of media. The plates were incubated for several weeks and the number of wells with cells growing were counted.

Monoclonal cells growing in these wells were transferred sequentially from the 96-well plate to a 24-well plate, 12-well plate, 6-well plate, T25, and T75 flask as they continued growing. The resulting monoclonal cells were used to plate a new 96-well plate in an attempt to select for the properties of cells that help them grow well as a single cell.

IV. Results

a. Sequencing

As shown in Figure 2, there were six TREs in LD with rs305088 that we designed primers for: U1, U2, U3, U4, U5, and U6 ("U" stands for upstream in reference to the TRE containing the rs305088 SNP). We sequenced the U1 region because it contained the rs305088 SNP and we wanted to confirm the SNP genotypes of the 23 CLL patients. We sequenced the U3 and U4 regions because the PRO-seq data indicated that these two regions had genotype-specific differences in the rates of transcription based on the rs305088 SNP genotype.

The U1 Sanger sequencing results for the 23 CLL patients confirmed the genotype information we received from our collaborators. No other SNPs were found in this TRE. As

shown in Table 3, the Sanger sequencing of U3 revealed four additional SNPs among the CLL patients, two of which closely followed the genotypic pattern of rs305088. The two SNPs located at chr16:85,936,495 and chr16:85,936,497 follow the same genotypic pattern as rs305088 except in patients 1N, CW233, and CW235. The SNP located at chr16:85,936,495 has been documented as rs2970089 and the SNP located at chr16:85,936,497 has been documented as rs2934499. As shown in Table 4, the Sanger sequencing of the U4 region revealed 16 additional SNPs, one of which closely followed the genotypic pattern of rs305088. This SNP is located at chr16:85,934,116 and follows the same genotypic pattern as rs305088 except in patients CW233 and CW235. The SNP located at chr16:85,934,116 has not been previously documented.

In order to have sequencing and genotypic information for the MEC1 and OSU-CLL cell lines, we performed Sanger sequencing in both these cell lines for all six rs305088 TREs. MEC1 and OSU-CLL cell lines have the risk allele for rs305088. Sanger sequencing of the region surrounding the rs4777184 SNP revealed a 5bp deletion from chr15:69,714,585-69,714,589 in 7 out of the 23 patient samples: 2N, JB17, CW22, CW219, CW230, CW233, and CW236. Sanger sequencing of this region also revealed a novel SNP at chr15:69,714,534.

Sanger sequencing of the region surrounding the rs674313 SNP proved to be very difficult because it is a previously documented highly polymorphic region and as a result we had trouble with the primers used for PCR and sequencing because the region is so highly variable (Shiina et al., 2009). We decided to focus on the region surrounding rs305088 because it was easier to study and we had located multiple SNPs in that region.

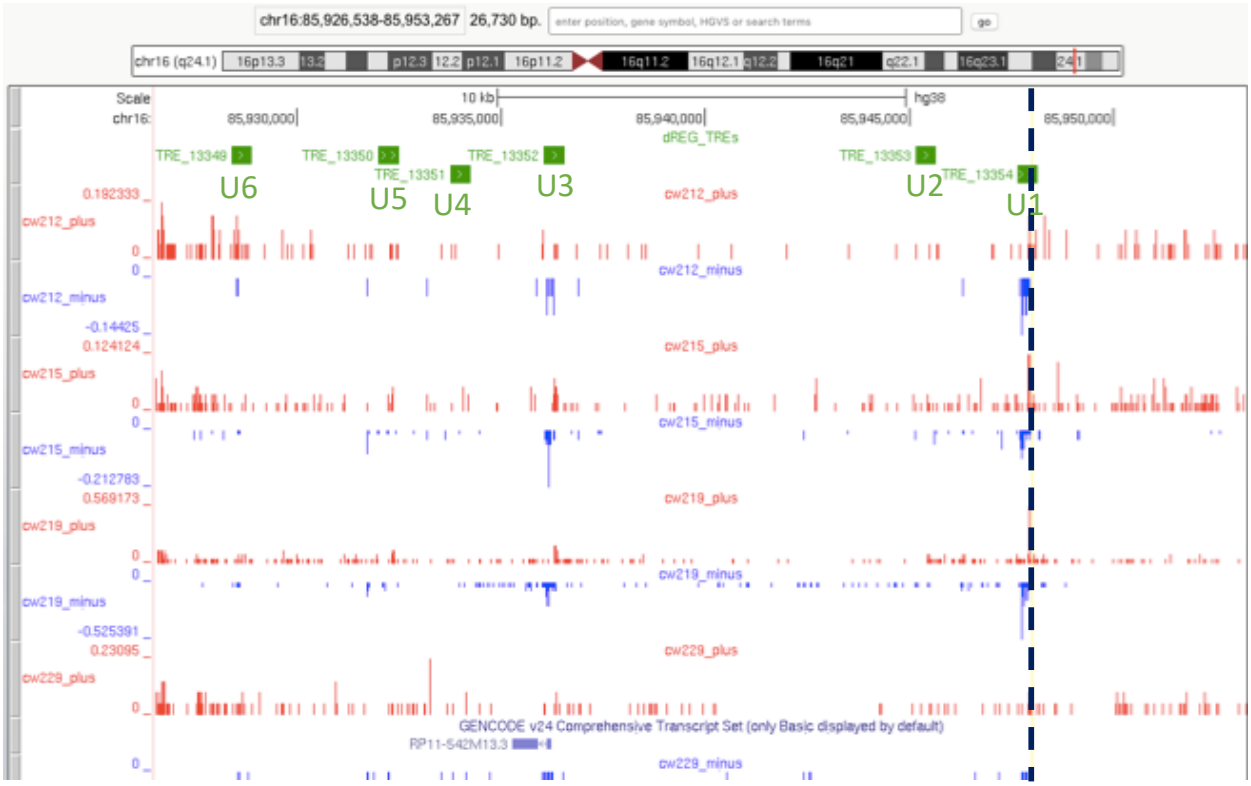


Figure 2: TREs in LD with rs305088. The dashed line indicates the location of rs305088 within U1.

SNPs - chr16 - rs305088U3					
Sample	chr16:85,936,132	chr16:85,936,446	chr16:85,936,495	chr16:85,936,497	rs305088
4N	AA	CC	AA	TT	AA
CW212	AA	CC	AA	TT	AA
CW215	AA	CC	AA	TT	AA
CW226	AA	CC	AA	TT	AA
1N	AA	CC	AA	TT	AG
2N	AA	CC	AT	CT	AG
5N	AA	CG	AT	CT	AG
CW189	AG	CC	AT	CT	AG
CW229	AG	CC	AT	CT	AG
CW236	AA	CC	AT	CT	AG
S19	AA	CC	AT	CT	AG
JB15	AA	CC	AT	CT	AG
3N	AA	CC	TT	CC	GG
CW188	AG	CC	TT	CC	GG
CW219	GG	CC	TT	CC	GG
CW22	AG	CC	TT	CC	GG
CW230	AG	CC	TT	CC	GG
CW231	AA	CG	TT	CC	GG
CW233	AG	CC	AT	CT	GG
CW235	AG	CC	AT	CT	GG
CW248	GG	CC	TT	CC	GG
JB17	AA	GG	TT	CC	GG
S20	AG	CG	TT	CC	GG
OSU	AA	CC	TT	CC	GG
MEC1	AA	CC	TT	CC	GG
Template	AA	CC	AA	TT	AA
Risk Allele			AA?	TT?	AA

Table 3: Genotypes of four SNPs identified with Sanger sequencing of the U3 region compared to the patient genotypes for the rs305088 SNP. The two SNPs located at chr16:85,936,495 and chr16:85,936,497 follow the same genotypic pattern as rs305088 in all patients except 1N, CW233, and CW235.

	SNPs - chr16 - rs305088U4																
Sample	chr16:85,933,769	chr16:85,933,902	chr16:85,933,988	chr16:85,933,998	chr16:85,934,034	chr16:85,934,051	chr16:85,934,075	chr16:85,934,083	chr16:85,934,116	chr16:85,934,129	chr16:85,934,138	chr16:85,934,159	chr16:85,934,272	chr16:85,934,280	chr16:85,934,285	chr16:85,934,314	rs305088
4N	TG	TT	GG	CC	AA	TT	CC	GG	CC	AA	TT	AA	CC	GG	GG	AA	AA
CW212	TG	TT	GG	CC	AA	TT	CC	GG	CC	AA	TT	AA	CC	GG	GG	AA	AA
CW215	TG	TT	GG	CC	AA	TT	CC	GG	CC	AA	TT	AA	CC	GG	GG	AA	AA
CW226	TT	TT	GG	CC	AA	TT	CC	GG	CC	AA	TT	AA	CC	GG	GG	AA	AA
1N	TG	TT	GG	CC	AA	TT	CC	GG	CG	AA	TT	AA	CC	GG	GG	AA	AG
2N	TG	TC	GG	CG	AG	TC	CT	GA	CG	AT	TC	AG	CT	GA	GA	AG	AG
5N	TG	TC	GG	CC	AG	TT	CC	GG	CG	AA	TT	AA	CC	GG	GG	AG	AG
CW189	TG	TT	AG	CC	AA	TT	CC	GG	CG	GA	TT	AA	CC	GG	GG	AA	AG
CW229	TG	TT	AG	CC	AA	TT	CC	GG	CG	GA	TT	AA	CC	GG	GG	AA	AG
CW236	TG	TC	GG	CG	AG	TC	CT	GA	CG	AT	TC	AG	CT	GA	GA	AG	AG
S19	TG	TC	GG	CG	AG	TC	TT	GA	CG	AT	TC	AG	CT	GA	GA	AG	AG
JB15	TT	TC	GG	CG	AG	TC	CT	GA	CG	AT	TC	AG	CT	GA	GA	AG	AG
3N	TT	CC	GG	GG	GG	CC	TT	AA	GG	TT	CC	GG	TT	AA	AA	AA	GG
CW188	TG	TC	AG	CG	AG	TC	CT	GA	GG	GA	TC	AG	CT	GA	GA	AG	GG
CW219	TG	TT	AA	CC	AA	TT	CC	GG	GG	GG	TT	AA	CC	GG	GG	AA	GG
CW22	TG	TC	AG	CG	AG	TC	CT	GA	GG	AT	TC	AG	CT	GA	GA	AA	GG
CW230	TG	TC	AG	CG	AG	TC	CT	GA	GG	GT	TC	AG	CT	GA	GA	AA	GG
CW231	TG	CC	GG	CG	GG	TC	CT	GA	GG	AT	TC	AG	CT	GA	AA	AG	GG
CW233	TT	TT	AG	CC	AA	TT	CC	GG	CG	GA	TT	AA	CC	GG	GG	AA	GG
CW235	TG	TT	AG	CC	AA	TT	CC	GG	CG	GA	TT	AA	CC	GG	GG	AA	GG
CW248	TG	TT	AA	CC	AA	TT	CC	GG	GG	GG	TT	AA	CC	GG	GG	AA	GG
JB17	TG	CC	GG	CC	AA	TT	CC	GG	GG	AA	TT	AA	CC	GG	AA	AA	GG
S20	TG	TC	AG	CC	AG	TT	CC	GG	GG	GA	TT	AA	CC	GG	GA	AA	GG
OSU	TT	CC	GG	GG	GG	CC	TT	AA	GG	TT	CC	GG	TT	AA	AA	AG	GG
MEC1	TT	CC	GG	GG	GG	CC	TT	AA	GG	TT	CC	GG	TT	AA	AA	AG	GG

Table 4: Genotypes of 16 SNPs identified with Sanger sequencing of the U4 region compared to the patient genotypes of the rs305088 SNP. The SNP located at chr16:85,934,116 follows the same genotypic pattern as rs305088 in all patients except CW233 and CW235.

b. Transcription Factor Binding Site Analysis

We performed TOMTOM TF binding site analysis on the GWAS-identified and newly identified SNPs. Since performing the initial TOMTOM analysis, the software has been updated and the new software produces very different TF results than originally obtained. Both the initial and the most recent TOMTOM analyses are included.

b1. Initial TOMTOM Transcription Factor Binding Site Analysis

The first SNP we analyzed was rs305088 located in the U1 TRE on chromosome 16. Figure 3 outlines the sequence given and the top three TF hits for the risk and the normal allele. GAAAAAGTGAA was used as the sequence with the normal allele, A, at the rs305088 SNP location. GAAAAGGTGAA was used as the sequence with the risk allele, G, at the rs305088 SNP location.

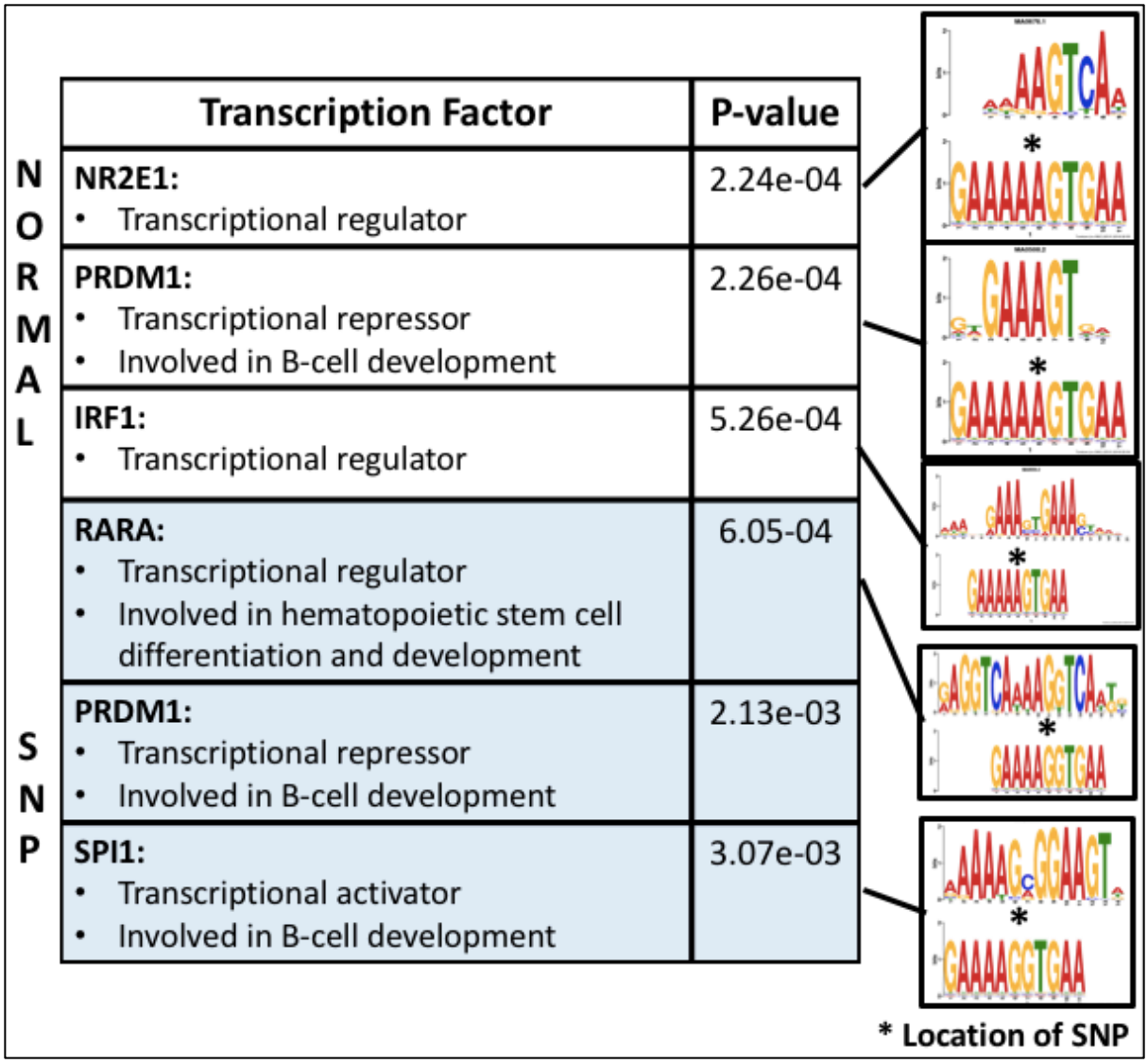


*Figure 3: Initial TOMTOM motif analysis top three hits for motif with normal allele at rs305088 SNP location and risk allele at rs305088 SNP location. * indicates the SNP location.*

The second and third SNPs we analyzed were located in the U3 TRE at chr16:85,936,495 and chr16:85,936,497. Figure 4 outlines the sequence given and the top three TF hits for the risk and the normal allele. Since these two SNPs are only one base pair apart, we compared the sequence with the risk allele at both SNPs to the sequence with the normal allele at both SNPs. GTCTATTTTTT was used as the sequence with the normal alleles, A and T at the SNP locations. GTCTTTCTTTT was used as the sequence with the risk alleles, C and G, at the SNP locations.

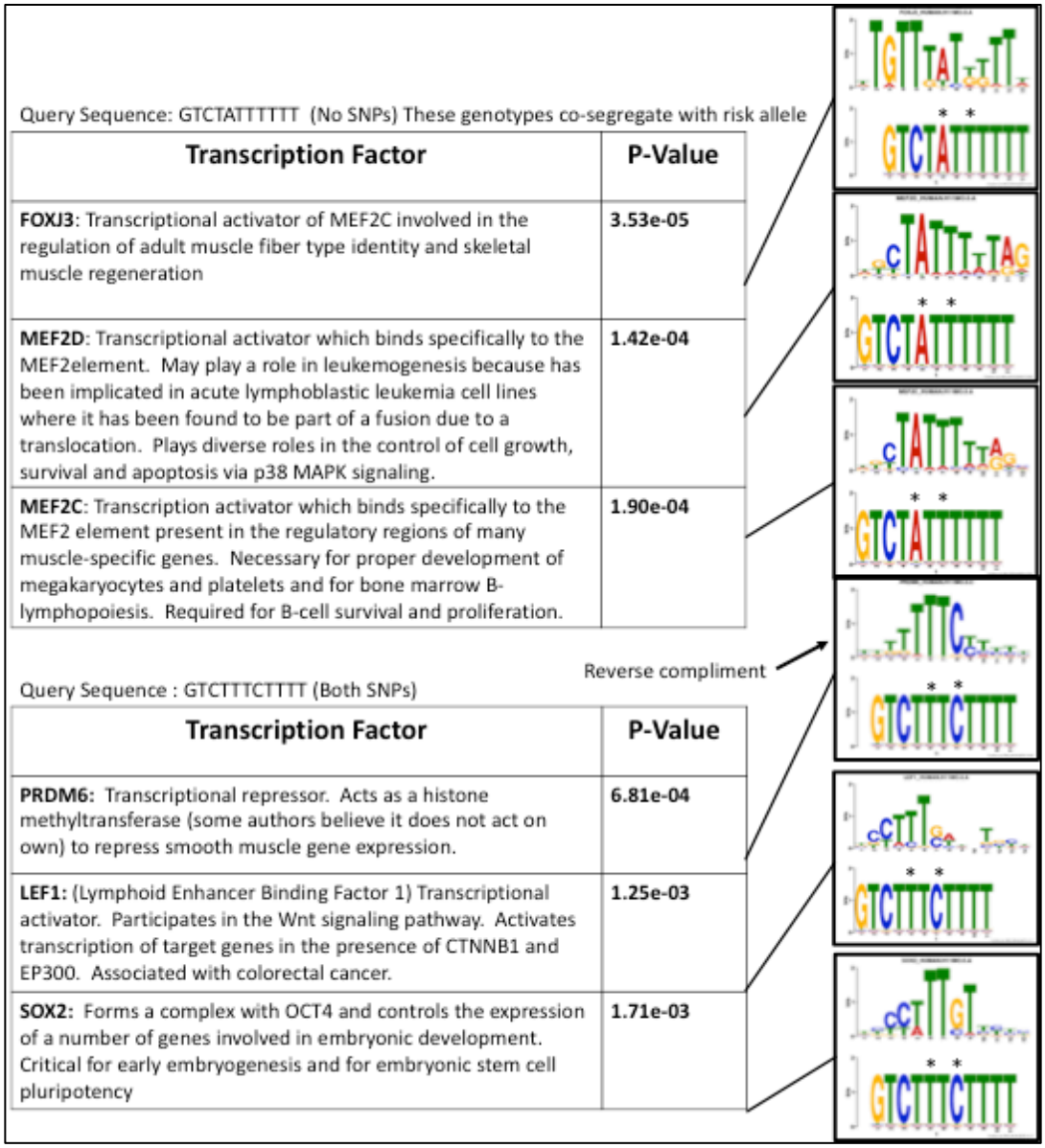


*Figure 4: Initial TOMTOM motif analysis top three hits for motif with normal alleles at the chr16:85,936,495 and chr16:85,936,497 SNP locations and risk alleles at the SNP locations. * indicates the SNP locations.*

The fourth SNP we analyzed was a new SNP that was identified at position chr16:85,934,116 which is located in the U4 TRE upstream of rs305088. Figure 5 outlines the sequence given and the top three TF hits for the risk and the normal allele. AGACTGCGGCTGG was used as the sequence with the normal allele, C, at the chr16:85,934,116 SNP location.

AGACTGGGGCTGG was used as the sequence with the risk allele, G, at the chr16:85,934,116 SNP location.

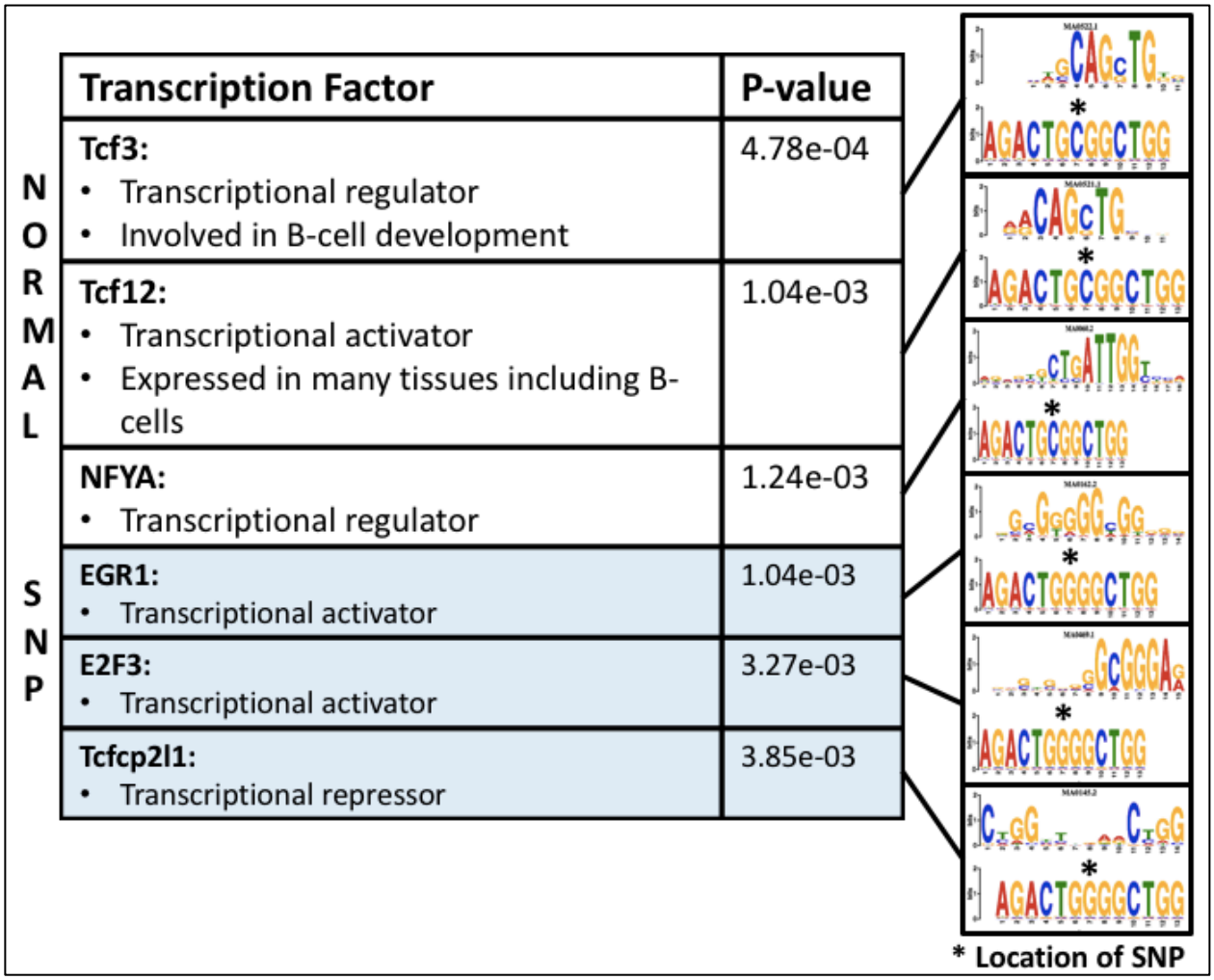


*Figure 5: Initial TOMTOM motif analysis top three hits for the motif with the normal allele and the risk allele at the chr16:85,934,116 SNP location. * indicates the SNP location.*

The fifth SNP we analyzed was found during sequencing of the rs4777184 region and is located at chr15:69,714,534. rs4777184 itself was not analyzed because it is not located in an active TRE region. Figure 6 outlines the sequence given and the top two TF hits for the normal allele. There were no TF hits for the sequence with the risk allele at the SNP position. TCAGTGGAAGA was used as the sequence with the normal allele, G, at the rs4777184 SNP

location and TCAGTAGAAGA was used as the sequence with the risk allele, A, at the SNP location.

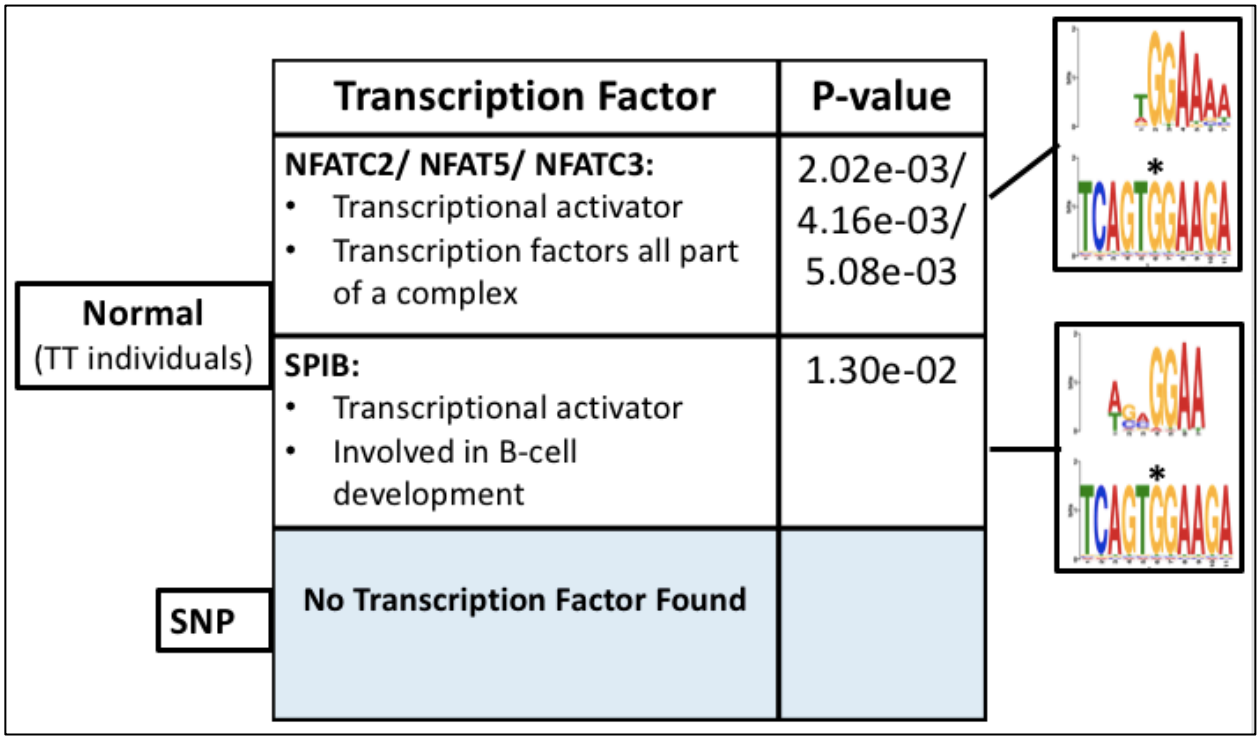


*Figure 6: Initial TOMTOM motif analysis top hits for motif with normal allele and the risk allele at the chr15:69,714,534 SNP location. * indicates the SNP location.*

During sequencing of the rs4777184 region, a 5bp deletion was found in the region chr15:69,714,585-69,714,589. Figure 7 outlines the sequence given and the top TF hits for the sequence with the deletion and with no deletion. GGGGCTTCTTG was used as the sequence without the deletion and GAGGGTTGCG was used as the sequence with the 5bp deletion.

In total, initial TOMTOM motif analysis revealed 26 TFs that are predicted to either gain or lose a TF binding site due to the presence of the SNPs in the TREs surrounding rs305088 and rs4777184. These include: E2F3, EGR1, FOXH1, FOXJ3, IRF1, LEF1, MEF2C, MEF2D, NFAT5, NFATC2, NFATC3, NFYA, NR2E1, PRDM1, PRDM6, PROX1, RARA, SOX2, SPI1, SPIB, SPZ1, Tcf12, Tcf3, Tcfcp2l1, TEAD4, and YY1.

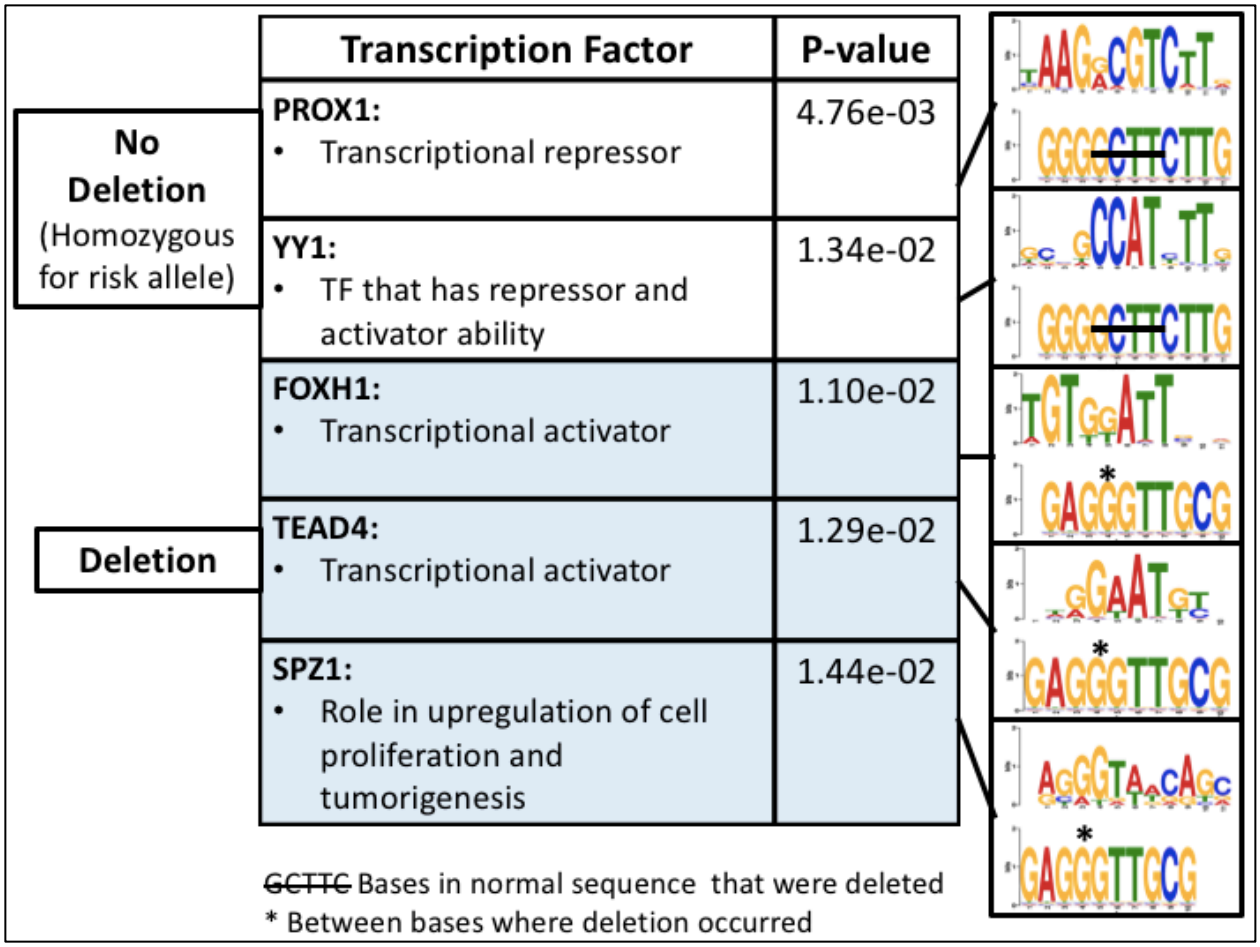


Figure 7: Initial TOMTOM motif analysis top hits for motif with the normal sequence near rs4777184 and the sequence with the 5bp deletion of GCTTC.

b2. Most Recent TOMTOM Transcription Factor Binding Site Analysis

The first SNP analyzed was rs305088 located in the U1 TRE on chromosome 16. Figure 8 outlines the sequence given and the top three TF hits for the risk and the normal allele. CTTCACATTTTCA was used as the sequence with the normal allele, A, at the rs305088 SNP location. CTTCACGTTTTCA was used as the sequence with the risk allele, G, at the rs305088 SNP location.

The second and third SNPs analyzed were located in the U3 TRE at chr16:85,936,495 and chr16:85,936,497. Figure 9 outlines the sequence given and the top three TF hits for the risk and the normal allele. Since these two SNPs are only one base pair apart, we compared the sequence

with the risk allele at both SNPs to the sequence with the normal allele at both SNPs. TTGTCTATTTTTTTT was used as the sequence with the normal alleles, A and T, at the SNP locations. TTGTCTCTGTTTTTT was used as the sequence with the risk alleles, C and G, at the SNP locations.

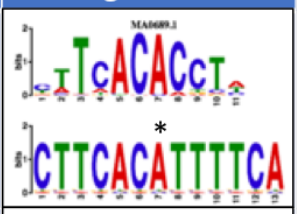
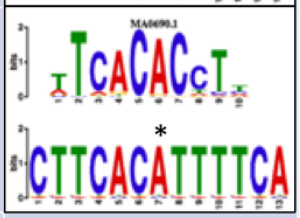
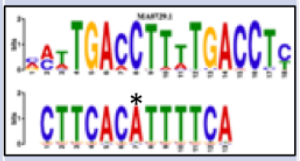
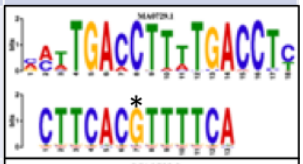
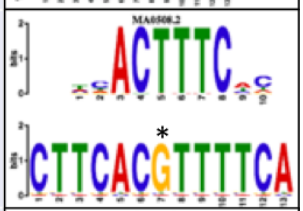
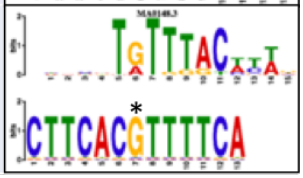
rs305088	Transcription Factor	P-value	Motif Position-Weight Matrix
Normal Allele (A)	TBX20 - Transcriptional activator and repressor for cardiac development	1.42e-03	
	TBX21 - Acts as a regulator of antiviral B-cell responses; initiates Th1 lineage development from naive Th precursor cells	1.45e-03	
	RARA - Implicated in regulation of development, differentiation, apoptosis, granulopoeisis, and transcription of clock genes. Translocations between this locus and several other loci have been associated with acute promyelocytic leukemia.	2.44e-03	
Risk Allele (G)	RARA - See above	1.80e-03	
	PRDM1 - Drives the maturation of B-lymphocytes into Ig secreting cells; repressor of beta-interferon gene expression	5.03e-03	
	FOXA1 - Transcriptional activator for liver-specific transcripts	8.20e-03	

*Figure 8: Most recent TOMTOM motif analysis top three hits for motif with normal allele at rs305088 SNP location and risk allele at rs305088 SNP location. * indicates the SNP location.*

The fourth SNP analyzed was a new SNP that was identified at position chr16:85,934,116 which is located in the U4 TRE upstream of rs305088. Figure 10 outlines the sequence given and the top three TF hits for the risk and the normal allele. AGACTGCGGCTGG was used as the sequence with the normal allele, C, at the chr16:85,934,116 SNP location. AGACTGGGGCTGG was used as the sequence with the risk allele, G, at the chr16:85,934,116 SNP location.

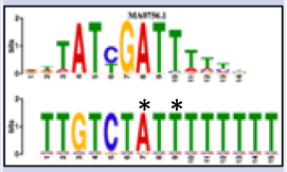
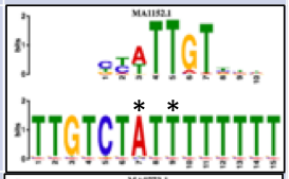
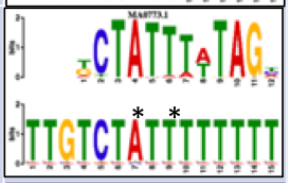
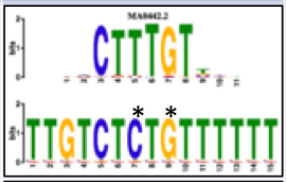
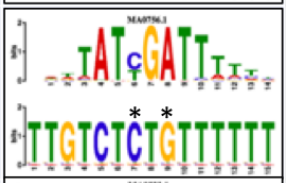
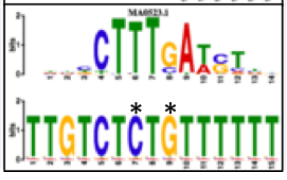
rs305088_U3	Transcription Factor	P-value	Motif Position-Weight Matrix
Normal Alleles (A, T)	ONECUT1/ONECUT2/ONECUT3 – Transcriptional activator; stimulates expression of target genes, including genes involved in melanocyte and hepatocyte differentiation	2.54e-04	
	SOX15 - Transcriptional regulator; involved in the regulation of embryonic development and in the determination of the cell fate	9.98e-04	
	MEF2D – Transcriptional activator; associated with lymphoblastic leukemia	2.47e-03	
Risk Alleles (C, G)	SOX10 - Involved in the regulation of embryonic development and in the determination of the cell fate	6.17e-04	
	ONECUT2 – See above	5.81e-03	
	TCF7L2 – Involved in in blood glucose homeostasis	6.72e-03	

*Figure 9: Most recent TOMTOM motif analysis top three hits for motif with normal alleles at the chr16:85,936,495 and chr16:85,936,497 SNP locations and risk alleles at the SNP locations. * indicates the SNP locations.*

The fifth SNP analyzed was found during sequencing of the region surrounding rs4777184 and is located at chr15:69,714,534. Figure 11 outlines the sequence given and the top three TF hits for the normal allele. There were no TF hits for the sequence with the risk allele at the SNP position. CTCAGTGGAAGAC was used as the sequence with the normal allele, G, at the rs4777184 SNP location and CTCAGTAGAAGAC was used as the sequence with the risk allele, A, at the SNP location.

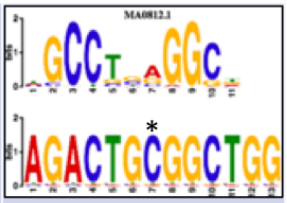
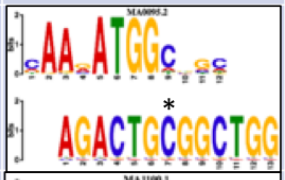
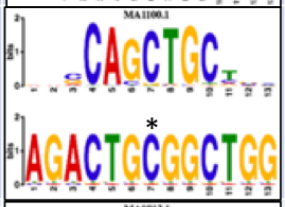
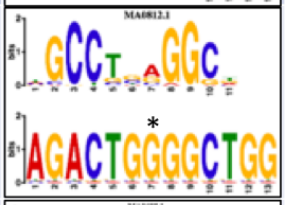
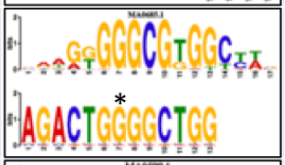
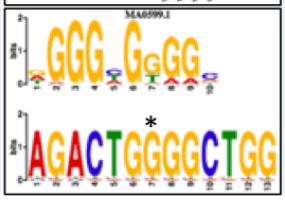
rs305088_U4	Transcription Factor	P-value	Motif Position-Weight Matrix
Normal Allele (C)	TFAP2A/TFAP2B/TFAP2C - Stimulates cell proliferation and suppress terminal differentiation of specific cell types during embryonic development	1.26e-03	
	YY1 – Has both activator and repressor activity	5.05e-03	
	ASCL1 – Transcriptional activator involved in neuronal commitment and differentiation	6.19e-03	
Risk Allele (G)	TFAP2A/TFAP2B/TFAP2C – See above	6.21e-04	
	SP4 – Transcriptional activator	4.06e-03	
	KLF5 – Involved in both promoting and suppressing cell proliferation	5.68e-03	

*Figure 10: Most recent TOMTOM motif analysis top three hits for the motif with the normal allele and the risk allele at the chr16:85,934,116 SNP location. * indicates the SNP location.*

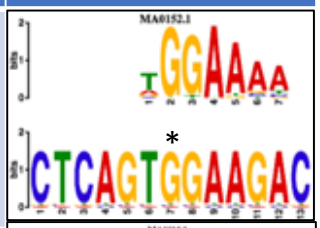
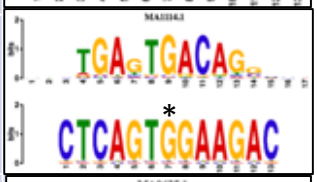
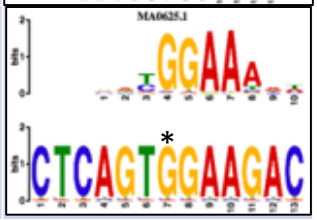
rs4777184 New SNP	Transcription Factor	P-value	Motif Position-Weight Matrix
Normal Allele (G)	NFATC2 - Plays a central role in inducing gene transcription during the immune response	2.64e-03	
	PBX3 (Pre-B-Cell Leukemia Transcription Factor 3) – Transcriptional activator	9.38e-03	
	NFATC3 - Regulator of transcriptional activation; involved in the immune system	9.73e-03	
Risk Allele (A)	No Results	N/A	N/A

*Figure 11: Most recent TOMTOM motif analysis top hits for motif with normal allele and the risk allele at the chr15:69,714,534 SNP location. * indicates the SNP location.*

During sequencing of the rs4777184 region, a 5bp deletion was found in the region from chr15:69,714,585-69,714,589. Figure 12 outlines the sequence given and the top TF hits for the sequence with the deletion and with no deletion. GAGGGGCTTCTTGCG was used as the sequence without the deletion and GAGGGTTGCG was used as the sequence with the 5bp deletion.

In total, TOMTOM motif analysis revealed 25 TFs that are predicted to either gain or lose a binding site due to the presence of the SNPs in the TRE regions surrounding rs305088 and rs4777184. These include: ASCL1, FOXA1, FOXH1, KLF5, MEF2D, MYB, NFATC2,

NFATC3, ONECUT1, ONECUT2, ONECUT3, PBX3, PRDM1, RARA, SOX10, SOX15, SP4, TBX20, TBX21, TCF7L2, TEAD4, TFAP2A, TFAP2B, TFAP2C, and YY1.

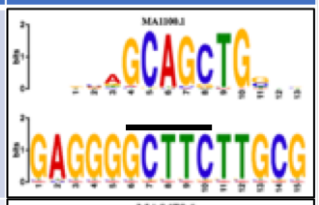
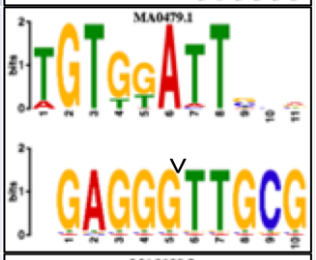
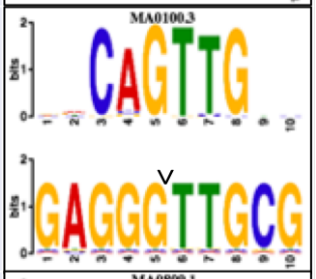
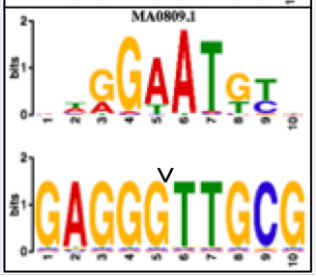
rs4777184 5bp deletion	Transcription Factor	P-value	Motif Position-Weight Matrix
No Deletion	ASCL1 – Transcriptional activator involved in neuronal commitment and differentiation	6.70e-03	
5bp Deletion	FOXH1 – Transcriptional activator	7.71e-03	
	MYB – Oncogene; may be aberrantly expressed or rearranged or undergo translocation in leukemias and lymphomas	8.54e-03	
	TEAD4 - Plays a key role in the Hippo signaling pathway, a pathway involved in organ size control and tumor suppression by restricting proliferation and promoting apoptosis	9.63e-03	

Figure 12: Most recent TOMTOM motif analysis top hits for motif with the normal sequence near rs4777184 and the sequence with the 5bp deletion of GCTTC.

c. ChIP-seq Analysis

Table 5 outlines the number of uniquely mapping ChIP-seq reads to the hg38 genome, the number of peaks called by MACS2 and the number of high confidence peaks between the two replicates of ChIP-seq data.

Dataset	Sequencing Depth	Number of Peaks	Number of High Confidence Peaks	Number of High Confidence Summits
Gm12878-NFYA-rep1	11,383,438	615	646	1,211
Gm12878-NFYA-rep2	11,943,138	1,548		
Gm12878-Tcf12-Rep1	14,051,006	59,417	45,900	66,000
Gm12878-Tcf12-Rep2	13,629,919	36,767		
Gm12878-Tcf3-Rep1	6,868,500	38,524	43,145	61,465
Gm12878-Tcf3-Rep2	15,242,225	54,067		
K562-Egr1-Rep1	18,848,323	53,551	42,668	63,577
K562-Egr1-Rep2	27,984,930	33,360		
K562-IFNa6h-IRF1-rep1	23,565,682	4,859	1,703	3,293
K562-IFNa6h-IRF1-rep2	22,938,815	1,668		
K562-IFNg30-IRF1-rep1	18,645,852	3,108	2,344	4,471
K562-IFNg30-IRF1-rep2	23,793,998	2,924		
K562-Sp1-Rep1	13,887,722	2,262	2,116	3,887
K562-Sp1-Rep2	17,252,822	5,743		
K562-Tead4-Rep1	27,988,989	49,721	69,457	92,805
K562-Tead4-Rep2	32,993,186	99,077		
K562-Yy1-Rep1	10,803,175	53,579	19,687	26,700
K562-Yy1-Rep2	6,784,368	22,951		

Table 5: ChIP-seq alignment and MACS2 results. Sequencing depth refers to the number of uniquely mapping reads.

Table 6 outlines the results from MAST before normalization. The total number of peaks refers to the number of sequences in the FASTA file that MAST searched to find matches to the TF motif. MAST was run twice for each TF – once using a TF motif that contained the normal allele for the SNP and a second time using a TF motif that contained the risk allele for the SNP. An E value of 10 was used as the cut-off for what we considered to be significant sequences that matched the motif. The "percent of sequences with E value <10" is the percentage of significant motifs out of the total number of peaks interrogated. Figure 13 shows the ChIP-seq results before normalization. The barplots represent the number of sequences with an E value < 10 for the sequence with the normal allele versus the number of sequences with an E value < 10 for the sequence with the risk allele.

Table 7 outlines the results from MAST after normalization. Normalization was performed by doing the same ChIP-seq analysis on all the TRE regions that were identified by dREG. The coordinates of the TRE regions were extracted, the center of each region found, and the windows extended 50bp upstream and 50bp downstream to get 101bp regions. The number of sequences observed in the ChIP-seq data divided by the number of sequences observed in the TRE data gave us the observed/total number. For most of the TFs, there were less sequences found in the TRE regions than in the ChIP-seq data. Figure 14 shows the ChIP-seq results after normalization using the TRE data. The number of sequences observed when the full TRE region was used was also found in order to compare the data to the 101bp TRE region data.

	SNP	Total Number of Peaks	Number of Sequences with E value < 10	Percent of Sequences with E value < 10
Egr1	C (normal allele)	63,577	1,121	1.763
	G (risk allele)		10,459	16.451
IFNa6hIRF1	A (normal allele)	3,293	78	2.369
	G (risk allele)		90	2.733
IFNg30IRF1	A (normal allele)	4,471	39	0.872
	G (risk allele)		49	1.096
NFYA	C (normal allele)	1,211	475	39.224
	G (risk allele)		172	14.203
Spi1	A (normal allele)	3,887	6	0.154
	G (risk allele)		23	0.592
Tcf12	C (normal allele)	66,000	381	0.577
	G (risk allele)		197	0.298

Table 6: MAST results before normalization. E value refers to the expected number of sequences that would randomly match the motif as well as the sequence data provided in a database of the same size.

	SNP	Total Number of Peaks	Number of Sequences with E value < 10	Number of Sequences in all TREs (101bp)	Number of Sequences in all TREs (full)	Observed/Total (101bp)
Egr1	C (normal allele)	63,577	1,121	230	312	4.874
	G (risk allele)		10,459	1,169	1,728	8.947
IFNa6hIRF1	A (normal allele)	3,293	78	41	46	1.902
	G (risk allele)		90	15	14	6.000
IFNg30IRF1	A (normal allele)	4,471	39	41	46	0.951
	G (risk allele)		49	15	14	3.267
NFYA	C (normal allele)	1,211	475	99	93	4.798
	G (risk allele)		172	35	28	4.914
Spi1	A (normal allele)	3,887	6	101	77	0.059
	G (risk allele)		23	205	110	0.112
Tcf12	C (normal allele)	66,000	381	49	0	7.776
	G (risk allele)		197	27	0	7.296

Table 7: MAST results after normalization.

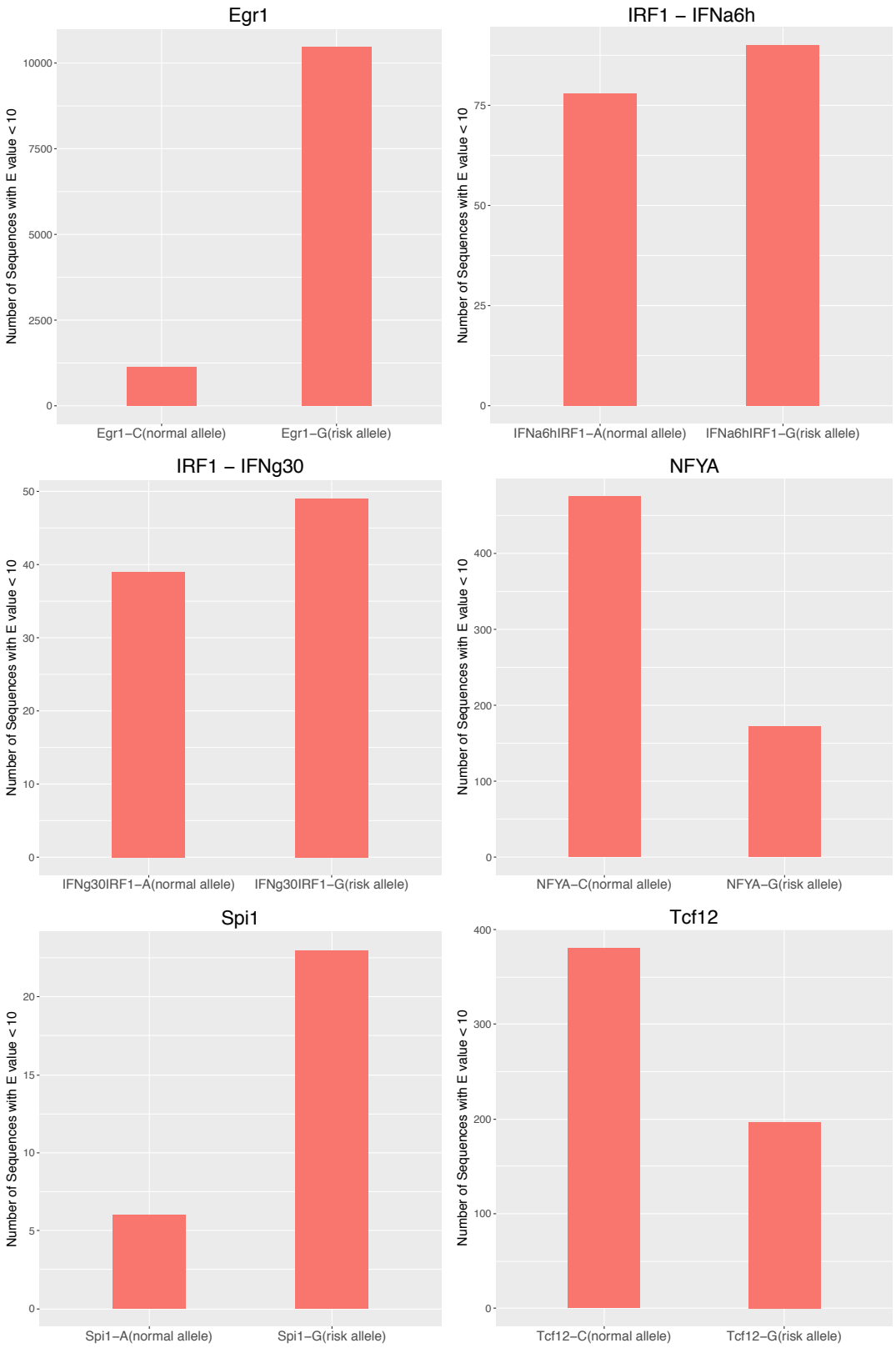


Figure 13: Barplots of ChIP-seq results before normalization.

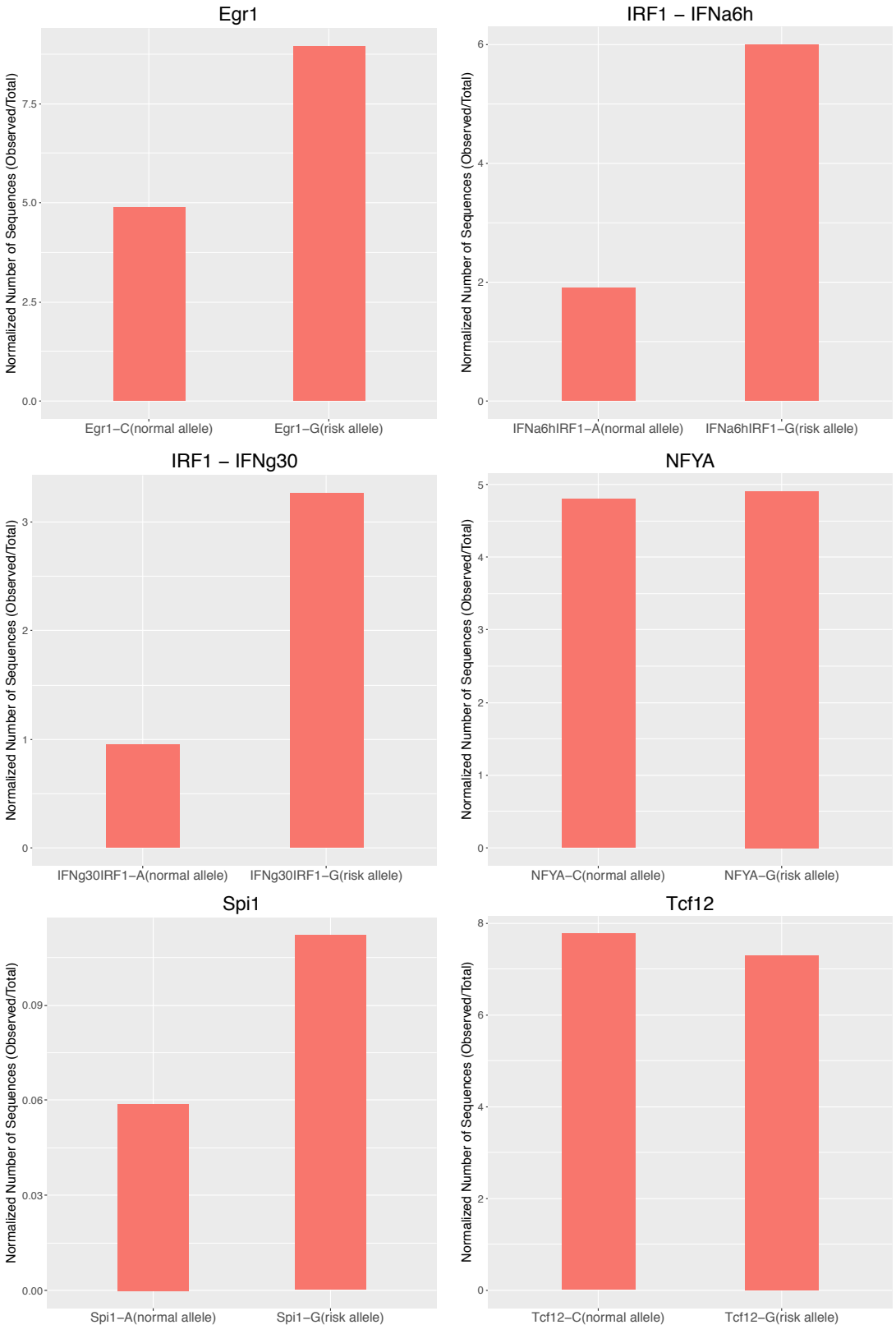


Figure 14: Barplots of ChIP-seq results after normalization.

d. Cell Culture

As shown in Table 8, The first MEC1 96-well plate was set up to have one cell in each well but resulted in only one well out of 96 growing monoclonal MEC1 cells. The cells were growing in a half conditioned, half new media well. When the well became confluent the cells were transferred sequentially to a 24-well plate, 12-well plate, 6-well plate, T25, and T75 flask. Some cells were frozen and others were used to start a new 96-well plate with these monoclonal cells in an attempt to select for cells that grow better in a single-celled environment. Two 96-well plates were set up with these monoclonal cells in 100% new media. The first monoclonal MEC1 plate had two wells out of 96 growing cells. The second monoclonal MEC1 plate had five wells out of 96 growing cells.

The second MEC1 96-well plate was set up to have two cells per well but only resulted in one well out of 96 growing monoclonal MEC1 cells. The cells were growing in a 100% new media well.

The OSU-CLL 96-well plate resulted in one well out of 96 growing monoclonal OSU-CLL cells. This well had 100% conditioned media. The cells were transferred to a 24-well plate but they did not continue growing.

Since each well with cells growing was from a different type of media, we concluded that the type of media did not make a difference and we used 100% new media from then on.

Cell Type	Concentration of Cells/Well	Number of Wells With Cells Growing	Media Type Growing Cells Found In
MEC1	1 cell/150 uL/well	1	half conditioned, half new media
MEC1	2 cells/100 uL/well	1	new media
Monoclonal MEC1	2 cells/100 uL/well	2	new media
Monoclonal MEC1	2 cells/100 uL/well	5	new media
OSU	1 cell/150 uL/well	1	conditioned media

Table 8: Growing Monoclonal MEC1 and OSU-CLL Cells.

V. Discussion

a. Sequencing

The discovery of many SNPs with genotypes that co-segregate with the CLL-associated GWAS-identified SNPs supports the data that these regions are in linkage disequilibrium and that there are many potentially causal SNPs in addition to the GWAS-identified CLL-associated SNPs. 1N, CW233, and CW235 were outliers in the co-segregation of the newly identified SNPs at chr16:85,936,495 and chr16:85,936,497 with the rs305088 patient genotypes. CW233, and CW235 were outliers in the co-segregation of the newly identified SNP at chr16:85,934,116 with the rs305088 patient genotypes. It remains unclear why these patient samples do not follow the genotypic patterns of the rest of the samples. The fact that the two CLL cell lines – MEC1 and OSU-CLL – have the risk allele at the rs305088 SNP reinforces the fact that these cell lines are a good model to use in further experiments and testing. rs305088 is located in an active TRE, however, rs4777184 is not located in an active TRE.

b. Transcription Factor Binding Site Analysis

In between the time that the initial TF binding site analysis was done and when the ChIP-seq analysis was finished, TOMTOM got updated and the updated version gives different TF results than were originally obtained. As a result, the TF binding site analyses for all of the SNPs was redone. However, all of the ChIP-seq analysis was done on TFs from the original TOMTOM binding site analysis. Therefore, most of the TFs that were analyzed with ChIP-seq data no longer show up as important TFs in the most recent TOMTOM analysis.

b1. Initial TOMTOM Transcription Factor Binding Site Analysis

Initial TF binding site analysis of rs305088 is invalid because the reverse complement sequence needs to be used in order to use G and A as the genotypes for the SNP.

Initial TF binding site analysis of the newly identified SNPs at chr16:85,936,495 and chr16:85,936,497 in the rs305088_U3 region suggests that FOXJ3, MEF2D, and MEF2C might lose their binding sites with the risk alleles at the chr16:85,936,495 and chr16:85,936,497 SNP locations. This may be important in the development of CLL because FOXJ3 is a transcriptional activator of MEF2C which is necessary for B cell survival and proliferation. PRDM6, LEF1, and SOX2 seem to gain a binding site with the risk alleles at the chr16:85,936,495 and chr16:85,936,497 SNP locations. It is unclear based on the annotations of these TFs how this may be involved in the development of CLL.

Initial TF binding site analysis of the newly identified SNP at chr16:85,934,116 in the rs305088_U4 region suggests that Tcf3, Tcf12, and NFYA might lose their binding sites with the risk allele at chr16:85,934,116. This may be important in the development of CLL because Tcf3 and Tcf12 are involved in the development of B cells. EGR1, E2F3, and Tcfcp2l1 seem to gain binding sites with the risk allele at chr16:85,934,116. More analysis needs to be done to determine how this might be related to the development of CLL.

Initial TF binding site analysis of the new SNP found at chr15:69,714,534 in the region surrounding rs4777184 suggests that NFATC2, NFAT5, NFATC3, and SPIB might lose their binding site with the risk allele at chr15:69,714,534. This may be important in the development of CLL because SPIB is involved in B cell development. TF binding site analysis also suggests that having the risk allele at chr15:69,714,534 destroys all TF binding sites for this region.

Initial TF binding site analysis of the 5bp deletion in the region nearby rs4777184, from chr15:69,714,585-69,714,589, suggests that PROX1 and YY1 lose their binding sites with the deletion. FOXH1, TEAD4, and SPZ1 seem to gain binding sites with the deletion. SPZ1 has been

documented to play a role in the upregulation of cell proliferation which may be related in some way to the development of CLL.

b2. Most Recent TOMTOM Transcription Factor Binding Site Analysis

Recent TF binding site analysis of rs305088 suggests that the binding site for RARA is not affected by the SNP genotype. The binding sites for TBX20 and FOXA1 seem to be affected by the SNP genotype but are likely unrelated to the CLL disease because these two TFs are involved in cardiac and liver development. The binding sites for TBX21 and PRDM1 also seem to be affected by the SNP genotype and might be important in the development of CLL because they are both involved in B cells. TBX21 seems to lose a binding site with the risk allele at the rs305088 SNP location and PRDM1 seems to gain a binding site with the risk allele at the rs305088 SNP location.

Recent TF binding site analysis of the newly identified SNPs at chr16:85,936,495 and chr16:85,936,497 in the rs305088_U3 region suggests that the binding sites for ONECUT1, ONECUT2, and ONECUT3 are not affected by the genotype of these two SNPs. The binding sites for SOX15, SOX10, and TCF7L2 seem to be affected by the SNP genotype but are unlikely to be related to the CLL disease because these TFs are involved in embryonic development and blood glucose homeostasis. MEF2D seems to lose a binding site with the risk alleles at the chr16:85,936,495 and chr16:85,936,497 SNP locations which might be important in the development of CLL because MEF2D has been associated with lymphoblastic leukemia.

Recent TF binding site analysis of the newly identified SNP at chr16:85,934,116 in the rs305088_U4 region suggests that the binding sites for TFAP2A, TFAP2B, and TFAP2C are not affected by the genotype of this SNP. YY1 and ASCL1 seem to lose their binding sites with the risk allele at chr16:85,934,116. SP4 and KLF5 seem to gain a binding site with the risk allele at

chr16:85,934,116. It is hard to predict whether any of these four TFs have any relation to CLL because they seem to have both transcriptional activator and repressor activity. More research needs to be done on how and if these TFs are related to the development of CLL.

Recent TF binding site analysis of the new SNP found at chr15:69,714,534 in the region surrounding rs4777184 suggests that NFATC2, PBX3, and NFATC3 might lose their binding sites with the risk allele at the SNP. This is interesting because PBX3 is a transcriptional activator associated with leukemia. There were no TF results when the risk allele at the SNP, which suggests that the risk allele destroys all TF binding sites at this region.

Recent TF binding site analysis of the 5bp deletion in the region nearby rs4777184 from chr15:69,714,585-69,714,589, suggests that ASCL1 loses a binding site with the 5bp deletion which is interesting because ASCL1 also seems to lose a binding site with the risk allele at chr16:85,934,116. FOXH1, MYB, and TEAD4 seem to gain a binding site with the 5bp deletion. This is interesting because MYB has been documented as an oncogene, however TEAD4 had been implicated in tumor suppression so these results are a bit conflicting.

Despite the interesting results from the TF binding site analysis, this analysis is just a computation prediction of which TFs are most likely to bind based on how well the TF motif matches a sequence. Therefore, this analysis is limited by the prediction algorithms and can only be used as a suggestion for which TFs might bind to a sequence – we cannot say for sure that these TFs do or do not bind to any of these regions.

c. ChIP-seq Analysis

The ChIP-seq analysis was limited by the fact that not all of the TFs that we wanted to analyze on had publicly available data in B cell lines. ChIP-seq analysis of the TFs supported the TF binding site analyses, however, there was no statistically significant difference in the number

of sequences that matched the TF binding motif with the normal allele versus with the risk allele at the SNP.

The barplots of the ChIP-seq data look very different before and after normalization. This indicates that normalization is very important to the proper analysis of this data because we are only interested in the TF binding sites within actively transcribed enhancers. Since for a majority of the TFs there were more sequences found in the ChIP-seq data than in the TRE data, this suggests that a lot of the sequences we are finding in the ChIP-seq data are part of inactive enhancers.

d. Cell Culture

The results from the experiments that aimed to grow monoclonal MEC1 and OSU-CLL cells indicates that these cells do not fare well living as single cells. This made it very difficult to establish monoclonal cells lines especially from the OSU-CLL cells.

VI. Future Directions

More analysis is needed to determine whether the TFs that were a result of the TOMTOM TF binding site analysis are actively transcribed in the patients we are investigating. If some of these TFs are not actively transcribed they are no longer of interest because we are looking for functionally active TFs. A closer examination of the software for TOMTOM also needs to be done to try to determine the source of the large discrepancies in the TF results between the two versions.

In the future, a luciferase reporter assay could be used to study the effect of select SNPs on enhancer activity. This would consist of a plasmid with a multiple cloning site where we would insert an enhancer of interest. We would insert the enhancer with and without the SNP to compare the difference that makes on transcription. ChIP-seq experiments can be performed to investigate

binding of TFs of interest in MEC1 and OSU-CLL cells lines. This will allow us to investigate all our TFs of interest and not be limited by the types of datasets that are publicly available.

Additionally, CRISPR and PRO-seq could be performed on two CLL cell lines (OSU-CLL and MEC1) to introduce SNPs of interest and to observe the effects of these SNPs on transcription, respectively. Lastly, we will want to confirm that the "novel SNPs" we have identified are indeed SNPs and not just common mutations among our patients. To do this we can sequence germline DNA from the saliva of patients to ensure that the germline DNA also contains these variants and that they are indeed SNPs and not somatic mutations.

VII. References

Bailey, Timothy L., and Michael Gribskov. "Combining evidence using p-values: application to sequence homology searches." *Bioinformatics (Oxford, England)* 14.1 (1998): 48-54.

Berndt, Sonja I., et al. "Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia." *Nature genetics* 45.8 (2013): 868-876.

Core, Leighton J., Joshua J. Waterfall, and John T. Lis. "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." *Science* 322.5909 (2008): 1845-1848.

Crowther-Swanepoel, Dalemari, et al. "Common variants at 2q37. 3, 8q24. 21, 15q21. 3 and 16q24. 1 influence chronic lymphocytic leukemia risk." Nature genetics 42.2 (2010): 132-136.

Danko, Charles G., et al. "Identification of active transcriptional regulatory elements from GRO-seq data." *Nature methods*12.5 (2015): 433.

Di Bernardo, Maria Chiara, et al. "A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia." Nature genetics 40.10 (2008): 1204-1210.

Ghia, Paolo, Andrés JM Ferreri, and Federico Caligaris-Cappio. "Chronic lymphocytic leukemia." *Critical reviews in oncology/hematology* 64.3 (2007): 234-246.

Goldin LR, Bjorkholm M, Kristinsson SY, Turesson I, Landgren O. “Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin’s lymphomas among relatives of patients with chronic lympocytic leukemia.” *Haematologica*. (2009); 94:647-53.

Gupta, Shobhit, et al. "Quantifying similarity between motifs." *Genome biology* 8.2 (2007): R24.

Kalil, Nelson, and Bruce D. Cheson. "Chronic lymphocytic leukemia." *The Oncologist* 4.5 (1999): 352-369.

Kwak, Hojoong, et al. "Precise maps of RNA polymerase reveal how promoters direct initiation and pausing." *Science*339.6122 (2013): 950-953.

Langmead, Ben, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome biology* 10.3 (2009): R25.

Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics*26.6 (2010): 841-842.

Sellick, G.S., Catovsky, D. and Houlston, R.S. “Familial chronic lymphocytic leukemia.” *Seminars in Oncology* (2006): 33, 195-201.

Shiina, Takashi, et al. "The HLA genomic loci map: expression, interaction, diversity and disease." *Journal of human genetics*54.1 (2009): 15.

Slager, Susan L., et al. "Genome-wide association study identifies a novel susceptibility locus at 6p21. 3 among familial CLL." Blood 117.6 (2011): 1911-1916.

Slager, Susan L., et al. "Genetic susceptibility to chronic lymphocytic leukemia." *Seminars in hematology*. Vol. 50. No. 4. WB Saunders, 2013.

Zhang, Yong, et al. "Model-based analysis of ChIP-Seq (MACS)." *Genome biology* 9.9 (2008): R137.